

Analyse lexicométrique d'un corpus multilingue de programmes électoraux

Ronny Scholz,

Université Paris XII

Université de Magdebourg (Allemagne)

Hypothèse et Questions

L'échelle de la ressemblance morphosyntaxique entre les partis politiques dépend de la ressemblance dans l'orientation politique de ces derniers, indépendamment du pays.

1. Quelle ressemblance morphosyntaxique existe entre les partis de même orientation politique mais de pays différents ?
 2. Y a-t-il une rhétorique commune aux partis de même orientation politique mais de pays différents ?
-

Démarche adoptée

- Présentation du corpus
 - Les partitions du corpus
 - Le logiciel TreeTagger - des avantages et des inconvénients
 - Problèmes d'unification des tags des différentes langues
 - Quelques expériences avec l'AFC du Lexico3
 - Conclusions
-

Caractéristiques du corpus

Principales caractéristiques lexicométriques

Nombre des occurrences	740077
Nombre des formes	20
Fréquence maximale	182869
Nombre des hapax	0
Nombre des textes	117

Ensemble des textes qui forment le corpus pour la période 1979 – 2004

corpus allemand:

REP*, CDU, CSU, FDP, Die Grünen, SPD, PDS*

corpus français:

FN, MPF, RPF, UMP (RPR, DL), UDF*, Les Verts, PS, PRG*, PCF, LO*

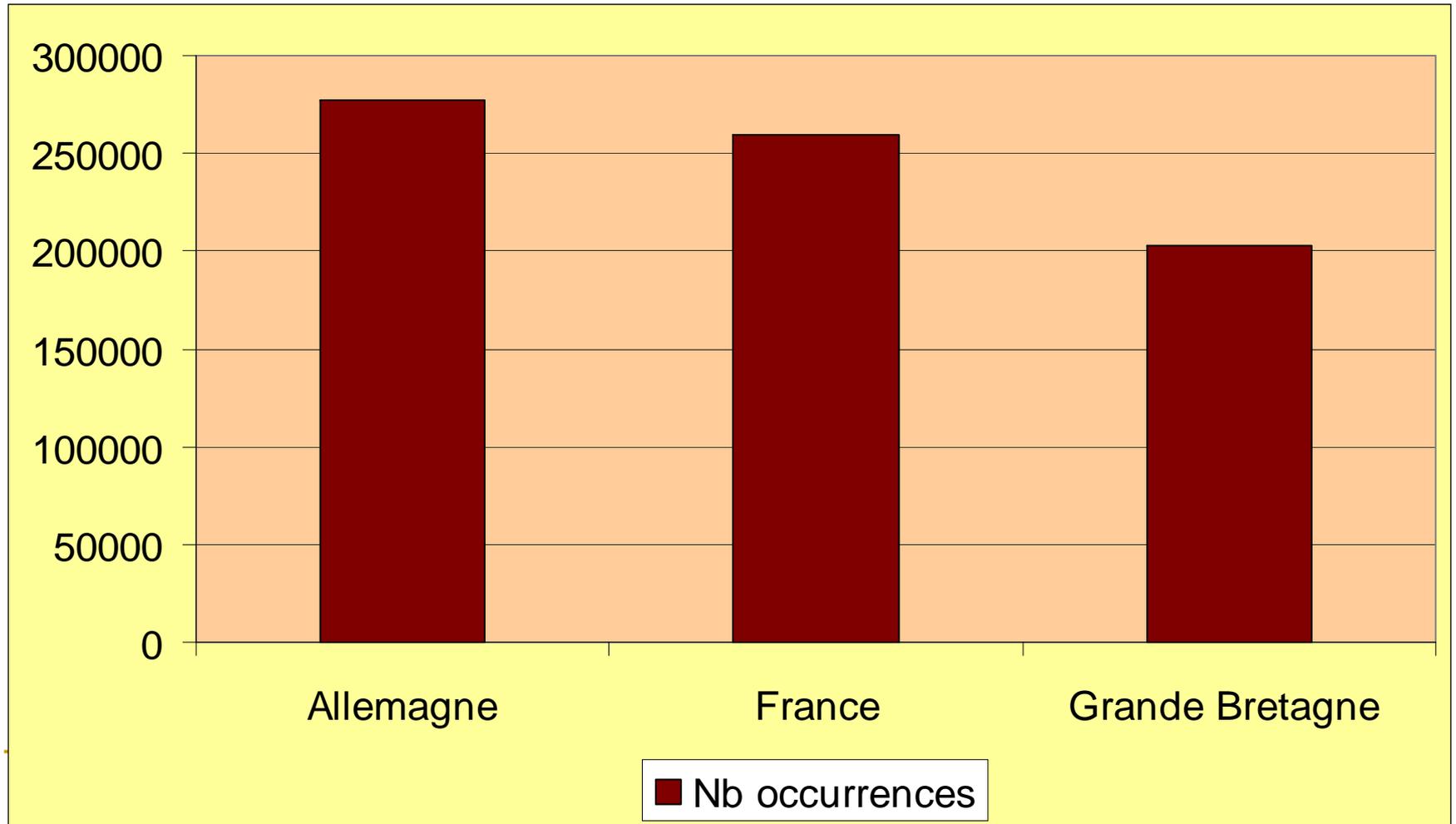
corpus britannique:

UKIP, CONS, LDP, LAB, GREENS, PC*, SNP*

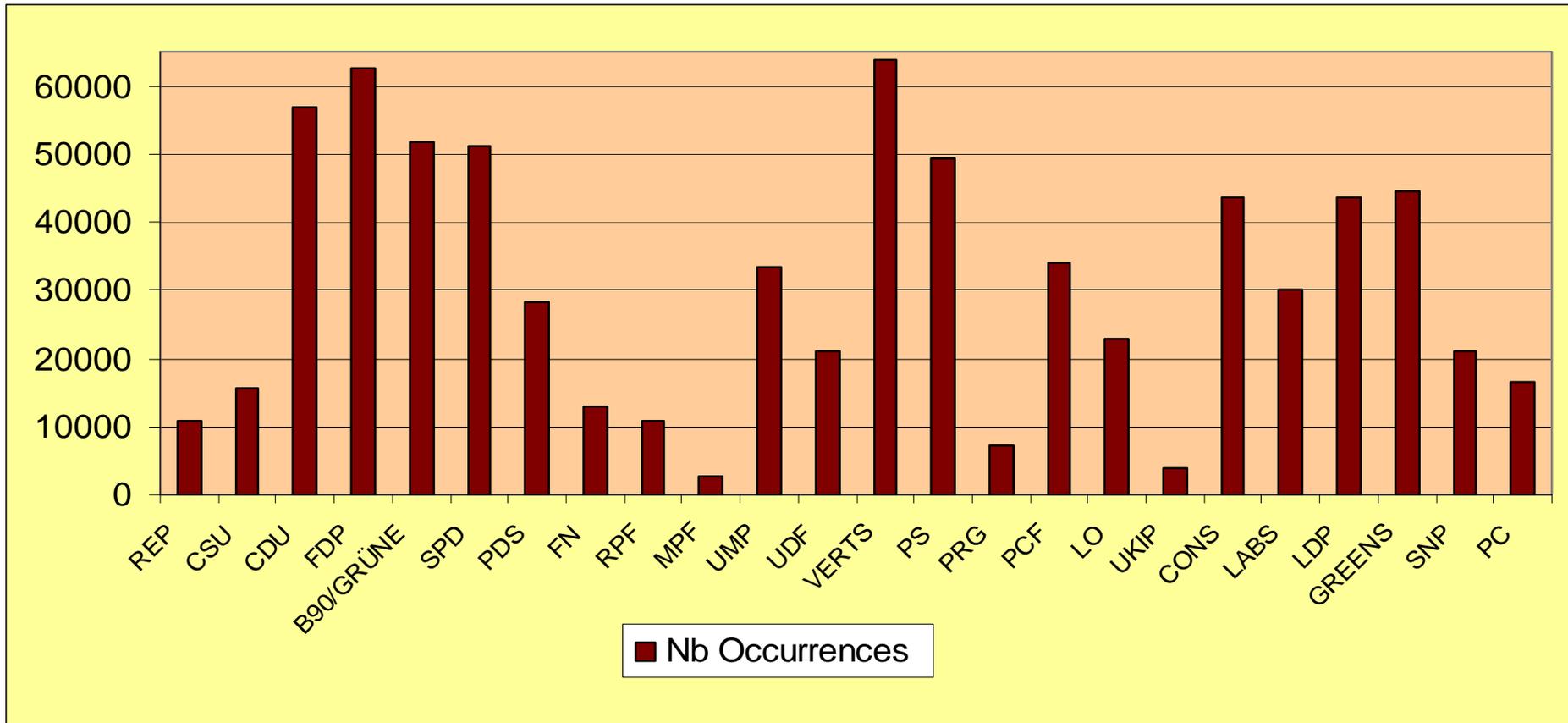
Partition du corpus :

1. Partition : `<jahr>` (année)
 - par exemple `<jahr=1979>`
 2. Partition : `<partei>` (parti)
 - par exemple `<partei=spd>`
 3. Partition : `<text1>` (texte)
 - par exemple `<text1=1979spd>`
 4. Partition : `<text2>` (texte)
 - par exemple `<text2=spd1979>`
 5. Partition : `<land>` (pays)
 - par exemple `<land=france>`
-

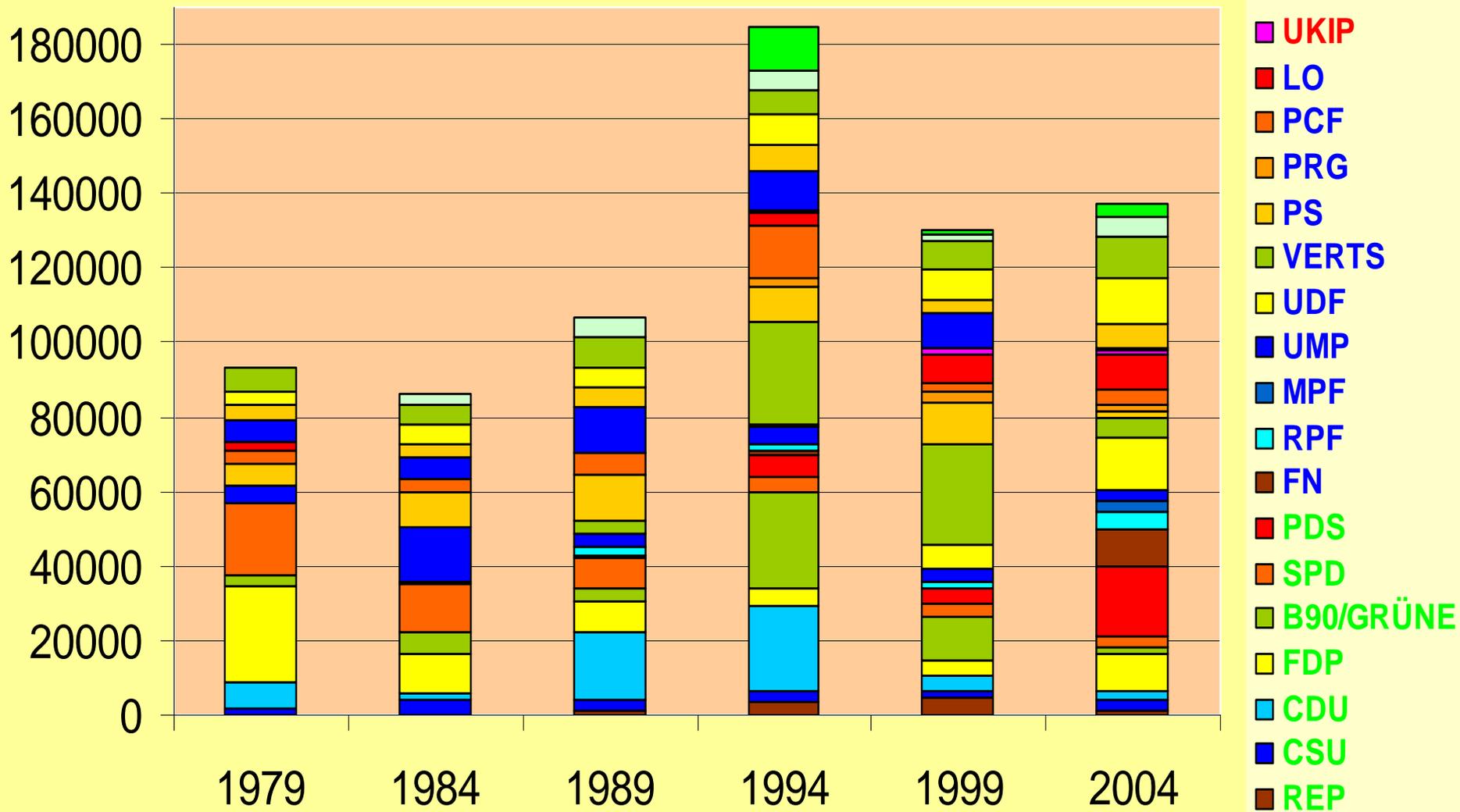
Distribution des fréquences absolues des textes de la partition <land> dans le corpus de programmes des Élections Européennes



Distribution des fréquences absolues des textes de la partition <parti> dans le corpus de programmes des Élections Européennes



Distribution des fréquences absolues des textes de la partition <jahr> (année) avec représentation des différents textes



Démarches pour former un corpus multilingue expérimental

- Tree-Tagger
 - Microsoft Excel
 - Unification des tags et ses problèmes
 - Lexico3
-

TreeTagger

TreeTagger (Windows)

Language

English

French

German

Italian

Spanish

Russian

Bulgarian

Dutch

Trained on

Latin-1

WindowsWE

Unicode

Task

Tag text I/O filenames pairwise in file

Output probability tree only

Output suffix/prefix tree only

Output for each token

the token

the speech-parts

best tag only

all tags with probability > times best tag

the tag probabilities

the lemma the token in place of unknown lemma

lexical information

none gramotron format

proto format proto format + probabilities

Suppress status messages

Input Options

Manual tags have probability

Manual tags have lemma

SGML tags present

End of sentence tag

Tokenization Options

none built-in own own + built-in

Own Program

Use abbreviations file

Use multi-word file

Auto-detach clitics

Auto-detach a word-final apostrophe

Convert Windows punctuation to Latin-1

1-letter word followed by period is abbreviation

Tagging Options

Use capital heuristics

Use hyphen heuristics

Use only lexical probabilities

Replace zero lexical frequencies by

Ignore prefix tree

Auxiliary lexicon

Input File

Output File

Tree Tagger developed by Helmut Schmid 
Version for MS Windows
Graph. Interface by Ciarán Ó Duibhín, 2008/03/09 

Corpus de vérification des tags

Microsoft Excel - 1989.xls

Frage hier eingeben

Arial 10

Bearbeitung zurücksenden... Bearbeitung beenden...

	A	B	C	D	E	F	G	H	I	J	K	L
1	Quatre	NUM	quatre									
2	raisons	NOM	raison									
3	de	PRP	de									
4	voter	VER:infi	voter									
5	Vert	NAM	<unknown>									
6	et	KON	et									
7	deux	NUM	deux									
8	interrogations	NOM	interrogation									
9	.	SENT	.									
10	La	DET:ART	le									
11	coordination	NOM	coordination									
12	générale	ADJ	général									
13	de	PRP	de									
14	la	DET:ART	le									
15	Nouvelle	NAM	<unknown>									
16	Gauche	NAM	<unknown>									
17	,	PUN	,									
18	le	DET:ART	le									
19	24	NUM	@card@									
20	avril	NOM	avril									
21	,	PUN	,									
22	s'est	NOM	<unknown>									
23	prononcée	VER:pper	prononcer									
24	pour	PRP	pour									
25	un	DET:ART	un									
26	soutien	NOM	soutien									
27	à	PRP	à									
28	la	DET:ART	le									
29	liste	NOM	liste									
30	«	PUN:cit	«									
31	Europe	NOM	<unknown>									

FN1989 / PCF_1989 / PRG1989 / PS_1989 / UDF1989 / RPRUDF_1989 / **verts1989**

Bereit

Corpus de vérification des tags

Microsoft Excel - 1989.xls

Datei Bearbeiten Ansicht Einfügen Format Extras Daten Fenster ? Adobe PDF

Frage hier eingeben

Arial 10

Bearbeitung zurücksenden... Bearbeitung beenden...

	A	B	C	D	E	F	G	H	I	J	K	L
1	NUM	Quatre	quatre									
2	NOM	raisons	raison									
3	PRP	de	de									
4	VER:infi	voter	voter									
5	NAM	Vert	<unknown>									
6	KON	et	et									
7	NUM	deux	deux									
8	NOM	interrogations	interrogation									
9	SENT	.	.									
10	DET:ART	La	le									
11	NOM	coordination	coordination									
12	ADJ	générale	général									
13	PRP	de	de									
14	DET:ART	la	le									
15	NAM	Nouvelle	<unknown>									
16	NAM	Gauche	<unknown>									
17	PUN	,	,									
18	DET:ART	le	le									
19	NUM		24 @card@									
20	NOM	avril	avril									
21	PUN	,	,									
22	NOM	s'est	<unknown>									
23	VER:pper	prononcée	prononcer									
24	PRP	pour	pour									
25	DET:ART	un	un									
26	NOM	soutien	soutien									
27	PRP	à	à									
28	DET:ART	la	le									
29	NOM	liste	liste									
30	PUN:cit	«	«									
31	NOM	Europe	<unknown>									

FN1989 / PCF_1989 / PRG1989 / PS_1989 / UDF1989 / RPRUDF_1989 / **verts1989**

Bereit

Corpus de vérification des tags

Lexico3 - [TextPloreur]

Fichier Traitement Fenêtre

Navigation Rapport Dictionnaire

Sélectionnez une couleur :

Recherche :

Formes (ordre lexicométrique)	Fréquence
pun_	17935
prp_de	14727
sent_	9821
detart_la	9322
detart_l	7379
kon_et	7180
prpdet_des	6918
detart_les	6757
prp_à	4953
detart_le	4720
prp_d	4081
prp_en	3279
detart_une	3225
prpdet_du	2690
detart_un	2462
prp_pour	2354
prp_dans	2279
verpres_est	2107
prorel_qui	1892
nam_europe	1774
kon_que	1726
prp_par	1713
prpdet_au	1670
pun_.	1665
adv_plus	1351
prp_sur	1336
-	1302
puncit	1302
proper_il	1260
proper_nous	1133
adj_européenne	1129
prpdet_aux	1060
adv_pas	1021
adv_ne	1018
prodem_ce	1018
nom_pays	959

18435 formes

Prêt

C:\Programme\Lexico 3\1979 - 2004 F ET_Forme Kleinbuchst.par

Feuille 1

<jahr=1979>
<partei=lo>
<text1=1979lo>
<text2=lo1979>
<land=frankreich>

_ " _puncit_" prp_pour detart_les nam_états-unis adj_socialistes nom_d' europe _ " _puncit_" nom_liste verpper_présentée prp_par nom_lutte adj_ouvrière kon_et detart_la nam_ligue nam_communiste nam_révolutionnaire pun_ (nom_section nom_francaise prp_de detart_la nam_quatrième nam_internationale pun_) sent_ . nom_liste nom_conduite prp_par nam_arlette nam_laguiller pun_ ; nam_alain nam_krivine sent_ . adj_travailleuses pun_ , nam_travailleurs pun_ , nam_1' assemblée nam_européenne kon_pour prorel_laquelle proper_nous verfutu_voterons detart_le num_10 nom_juin verpres_est verpper_présentée prp_par detpos_ses nom_partisans kon_comme detart_le nom_prélude prp_à nom_1' unification prp_de proind_tous detart_les nom_pays nom_d' europe sent_ . kon_mais proper_il nom_n' en verpres_est nom_rien sent_ . nam_1' assemblée nam_européenne versimp_n' aura proind_aucun nom_pouvoir adj_réal sent_ . detart_les nom_gouvernements kon_et detart_les adj_capitalistes nom_européens adv_ne verpres_sont adv_pas adv_près prp_de verinfi_supprimer detart_les adj_frontières prorel_qui verfutu_existeront adv_autant prp_après detart_le num_10 nom_juin adj_qu' avant sent_ . kon_et detpos_leur nam_marché nam_commun verpres_est verpper_construit prp_à nom_1' avantage prpdet_des adj_gros nom_industriels kon_et prpdet_des adj_multinationales sent_ . nam_c' est adv_pourquoi detart_les nom_travailleurs adj_manuels kon_et nom_intellectuels verfutu_n' ont nom_rien prp_à proper_en verinfi_attendre sent_ . adv_pourtant detpos_leur nom_intérêt verpres_est adv_bien verinfi_d' abolir prodem_ces adj_frontières prorel_qui verpres_ont verpper_été nom_1' objet prp_de adv_tant prp_de nom_guerres kon_et prorel_qui pun_ , adv_aujourd' hui pun_ , adv_ne verpres_servent adv_que prp_de nom_cadre prp_à detart_la nom_guerre adj_économique prorel_que proper_se verpres_font detart_les nom_trusts sent_ . kon_et adj_seuls detart_les nom_travailleurs verfutu_pourront adv_vraiment verinfi_unifier nom_1' europe pun_ , kon_car prodem_ce verpres_sont detart_les adj_seuls prp_à verinfi_pouvoir verinfi_créer detart_une adj_vaste nom_fédération prp_de nom_peuples nom_d' où verfutu_seront adv_définitivement adi_bannis nom_1' exploitation pun_ . detart_le nom_chômage kon_et detart_les

Sélectionnez une couleur : Recherche :

Formes (ordre lexicographique)	Fréquence
abr_i	5
abr_images	1
abr_insuffisances	1
abr_insulaires	1
abr_ivg	6
abr_km	9
abr_l	2
abr_livres	1
abr_mais	1
abr_monde	1
abr_n°	2
abr_nationalité	1
abr_nations	1
abr_natura	1
abr_nbc	1
abr_normes	1
abr_occidentale	1
abr_ocde	5
abr_on	1
abr_ong	1
abr_ont	1
abr_onu	33
abr_opa	3
abr_oser	1
abr_otan	3
abr_p	2
abr_paix	1
abr_pc	8
abr_pcf	14
abr_pci	1
abr_pdg	3
abr_pesticides	2
abr_pma	3
abr_pme	9
abr_pmi	1
abr_politique	1

18435 formes

 Forme : Tri : Regroupement : Largeur :
 Expression rationnelle Type de documentation : Délimiteurs de séquence :

```
om_nitrates pun_ , adj_herbicides pun_ , abr_pesticides . . sent_ . pun_ ) verpres_dé
adj_chimiques pun_ ( nom_engrais pun_ , abr_pesticides . . sent_ . pun_ ) kon_et det
```

Nombre de contextes : 2

Unification des tags - quelques cas douteux

Langue taggée:	française	anglaise	allemande	Synthèse
Article	DET:ART	DT	ART	ARTPROIND
Pronom indéfini	PRO:IND	DT PDT	PIS PIAT PIDAT	
Pronom possessif	DET:POS (ma, leur) PRO:POS (mien, vôtre)	PPS (POS) WPS	PPOSS PPOSAT	PROPOS
Pronom démonstratif	PRO:DEM	(PDT)	PDS PDAT	PRODEM
Mot étranger		(FW)	(FM)	effacé
Parti verbale séparé du verbe [er kommt] an			(PTKVZ)	effacé
Particule d'infinitif avec zu [Er begann] zu [laufen].			(PTKZU)	effacé

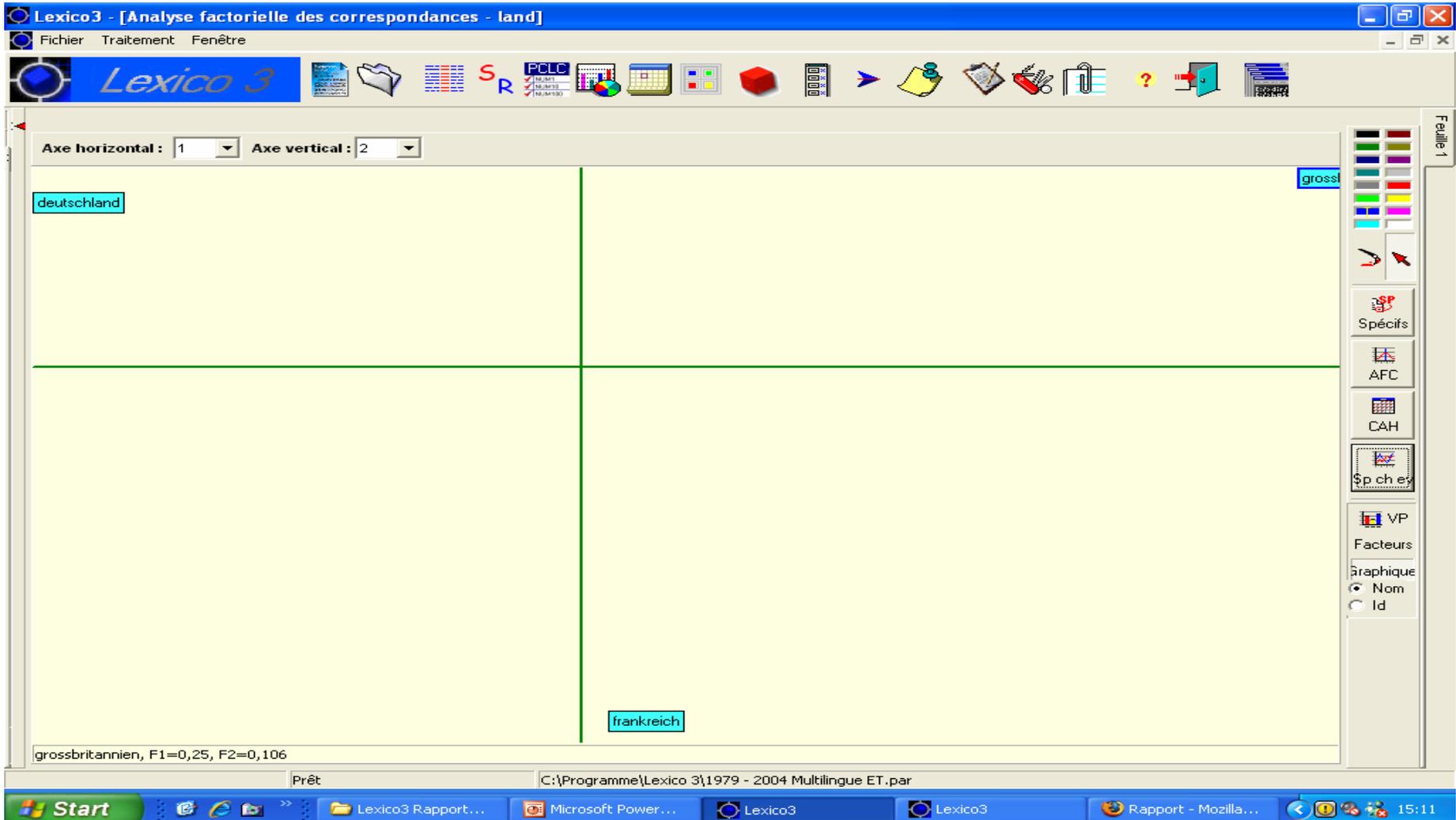


Verbe infinitif	VER:infi	VB (be) VH+ (have) VV	VVINF VVIZU VAINF VMINF	VERBINF
Verb présent	VER:pres VER:subp	VVP VVZ VBP VBZ VHP VHZ	VVFIN VAFIN VMFIN	VERBFIN
Verbe modal		MD		
Verbe futur	VER:futu			
Verbe passé	VER:impf VER:subi (imparfait) VER:simp	VBD VHD VVD		
Participe présent actif	VER:ppre	VBG VVG VHG	ADJD/ADJA	
Verbe impératif	VER:impe		VVIMP VAIMP	
Verbe conditional	VER:cond			

Index du corpus multilingue

Forme	Fréquence	Forme	Fréquence
NOM	182869	NUM	5916
PRÄP	110775	PRODEM	5843
ARTPROIND	107009	PRON	478
ADJ	83638	SYM	452
VERBFIN	54908	INT	74
KON	40095	,	40025
ADV	33800	.	35312
NAM	31941	:	3763
VERBINF	30698	(3367
PROPER	18077	”	757
PPP	17095	'	288
PROPOS	8391)	247
PROREL	7259		

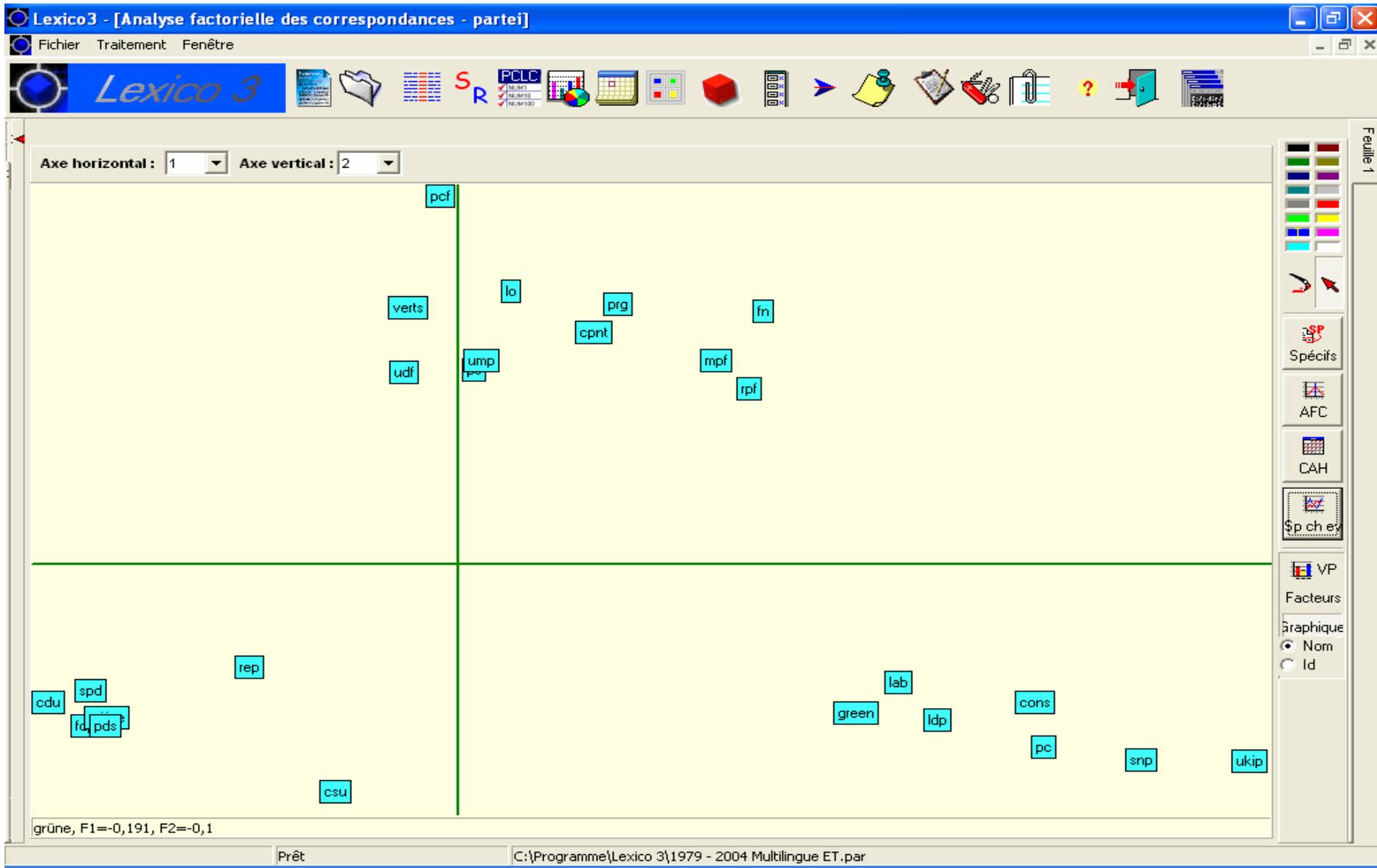
AFC de la partition <land> (pays)



Spécificités de la partition <land>

Formes	allemandes		français		britannique	
	<i>Fréquence</i>	<i>Spécificité</i>	<i>Fréquence</i>	<i>Spécificité</i>	<i>Fréquence</i>	<i>Spécificité</i>
NOM	74241	50	62901	-14	45727	-50
PRÄP	28364	-50	50387	50	32024	32
ARTPROIND	49374	50	35772	-35	21863	-50
ADJ	37420	50	24848	-50	21370	-41
VERBFIN	18988	-49	16836	-50	19084	50
KON	18159	50	12207	-50	9729	-50
ADV	12593	0	11863	0	9344	0
NAM	6226	-50	8459	-50	17256	50
VERBINF	11020	-9	9124	-50	10554	50
PROPER	6035	-31	7253	47	4789	-3
PPP	5824	-21	6124	2	5147	16
PROPOS	2772	-18	2885	0	2734	26
PROREL	2637	-2	3292	50	1330	-50
NUM	1196	-50	2879	50	1841	11
PRODEM	2222	0	3621	50	0	0
”	0	0	757	50	0	0
PRON	69	-29	88	-16	321	50
SYM	68	-26	286	34	98	-3
INT	24	0	39	3	11	-3

AFC de la partition <partei> (parti)



Spécificités de la partition <partei> des partis les plus extrêmes

Formes	CDU		CSU		PCF		UKIP	
	Fréquence	Spécificité	Fréquence	Spécificité	Fréquence	Spécificité	Fréquence	Spécificité
"		0		0	254	50		0
ADJ	7905	50	2132	22	3141	-36	329	-8
ADV	2200	-18		0	1349	-8	221	4
ARTPROIND	10566	50	2672	23	4619	-6	431	-9
INT		0	7	4	8	2		0
KON	3624	25	944	4	1592	-10	141	-7
NAM	955	-50	792	7	890	-50	438	50
NOM	15474	45	12207	-50		0	736	-16
NUM	337	-10	48	-15		0	51	4
PPP	1182	-5	302	-3		0	105	2
PRÄP	5754	-50	1549	-50	7035	50		0
PRODEM		0	328	-3	577	50		0
PRON	12	-7	85	-4	14	-2	9	3
PROPER	1184	-10	278	13	979	8		0
PROPOS	701	2	130	-2		0	78	6
PROREL	624	3	1	-4	387	3		0
SYM		0		0		0		0
VERBFIN	3642	-24		0	2020	-28	390	10
VERBINF		0	480	-12	1317	-3	197	3

Conclusions générales à partir de l'AFC

- Notre hypothèse:
« L'échelle de la ressemblance morphosyntaxique entre les partis politiques dépend de la ressemblance dans l'orientation politique de ces derniers, indépendamment du pays. » ne peut pas être confirmée. L'AFC montre bien que, au niveau morphosyntaxique, les partis se distinguent bien par rapport à la langue dans laquelle ils sont écrits. Mais sur le niveau morphosyntaxique il est possible de distinguer des partis « plus allemand », « plus français », « plus britannique » que les autres.
- Par rapport aux textes français et britanniques, la spécificité des textes **allemands** se situe au niveau des noms qui sont liés avec les articles, les adjectifs et les conjonctions.
- La spécificité des textes **français** se situe au niveau des différents pronoms. On peut dire que le discours français est pronominal, c'est-à-dire qu'il fonctionne sur des rapports entre les parties de phrase et les phrases mêmes. La grande spécificité des pronoms personnels indique un discours éthique dans le sens de la rhétorique aristotélicienne. La spécificité pronominale des textes français demande un plus grand effort de réflexion du lecteur que les textes anglais. Les pronoms créent en effet un espace virtuel tandis que les noms créent un espace plus direct et plus réel.
- La spécificité des textes **anglais** se situe au niveau des verbes conjugués et des verbes à l'infinitif, c'est-à-dire que le discours britannique suggère une plus grande activité que les textes français et allemand. Par ailleurs, les textes anglais contiennent beaucoup plus de noms propres que les textes allemands et français.

Les oppositions linguistiques ?

Noms:	+A	-F(-14)	-GB	= A ↔ GB
Pronoms personnels	-A	+F	-GB (-3)	= A ↔ F
Chiffres cardinaux	-A	+F	+GB (+11)	= A ↔ F
Pronoms relatifs	-A (-2)	+F	-GB	= F ↔ GB
Articles et pronoms indéfinis:	+A	-F	-GB	= A ↔ F et GB
Adjectifs:	+A	-F	-GB	= A ↔ F et GB
Conjonctions:	+A	-F	-GB	= A ↔ F et GB
Prépositions:	-A	+F	+GB	= A ↔ F et GB
Noms propres:	-A	-F	+GB	= A et F ↔ GB
Verbes conjugués:	-A	-F	+GB	= A et F ↔ GB
Verbes à l'infinitif	-A (-9)	-F	+GB	= A et F ↔ GB